DOCUMENT RESUME

ED 128 343 95 SP 010 416

AUTHOR Berliner, David C.
TITLE Impediments to the Study of Teacher Effectiveness.
INSTITUTION Far West Lab. for Educational Research and Development, San Francisco, Calif.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Nov 75
NOTE 34p.; Paper presented at the National Invitational Conference on Research on Teacher Effects: An Examination by Decision Makers and Researchers (Austin, Texas, November 3-5, 1975)
AVAILABLE FROM Far West Laboratory for Educational Research and Development, 1855 Folsom Street, San Francisco, California 94103 (No price quoted)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Academic Achievement; Classroom Research; *Educational Accountability; *Effective Teaching; Instrumentation; Learning Processes; *Performance Based Teacher Education; Predictive Ability (Testing); Research; Research Methodology; Standardized Tests; Statistical Analysis; Student Characteristics; Teacher Behavior; Teacher Education; Teacher Educators; *Teacher Evaluation; *Teacher Influence

ABSTRACT

The heart of performance- and competency-based teacher education, evaluation, and accountability programs is the establishment of empirical relationships between teacher behavior as an independent variable and student achievement as a dependent variable. Before researchers can adequately establish those relationships they need to deal with the problems of instrumentation, methodology, and statistics. Workers in this area must come to grips with the inadequacy of standardized tests, the unknown predictive validity of tests from special teaching units, the problem of building multivariate outcome measures, the problems of measurement of appropriateness of teacher behavior, the lack of experience in choosing an appropriate unit of analysis for describing teaching behavior, and the lack of stability of many teacher behaviors. Also discussed are the problems of how student background affects measures of teacher effectiveness, what subject matters should be examined, how normative standards and volunteer teachers affect what can be said about teachers and teaching, how individual students react to teaching skills, and how students monitor and interpret a teacher's behavior in ways that may or may not coincide with how educational theorists interpret the phenomena. (JMF)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

ED128343

# IMPEDIMENTS TO THE STUDY OF TEACHER EFFECTIVENESS

DAVID C. BERLINER

Far West Laboratory for Educational Research and Development

November 1975

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Paper Presented at the University of Texas, Austin, Texas, November 2-4, 1975, for the Conference on Research on Teacher Effects: An Examination by Decision-Makers and Researchers.

SP010 416

## PREFACE

The Far West Laboratory for Educational Research and Development has been conducting research on teacher effectiveness. Under contract to the California Commission for Teacher Preparation and Licensing, with funding provided by the National Institute of Education, the Laboratory is studying teachers in second- and fifth-grade classes in order to identify teacher behavior and classroom qualities that are related to reading and mathematics achievement.

The California Commission is the agency charged with certifying the appropriateness of teacher training programs throughout the state. To carry out its duties, the Commission needs information about what teacher behaviors are related to student outcomes. This information will then be used jointly by the Commission and the State institutions that it certifies in order to better insure that beginning teachers receive training in areas that have been empirically demonstrated to affect student learning.

To obtain the information they need, the Commission has undertaken a multi-year research effort entitled the Beginning Teacher Evaluation Study (BTES). During 1974-1975, as part of this study, the Laboratory did work on five major tasks. Three of these were substantive tasks, inquiring: 1) whether ethnographic approaches to the study of teaching could yield new insights into the teaching-learning process; 2) whether the perceptions of teachers who vary in their ability to induce learning differ when viewing teaching-learning situations, and, of equal interest, whether students' perceptions of the teaching-learning situation would yield information about what aspects of classroom interaction were salient and important to the students.

Two methodological issues were also addressed during 1974-1975. One of these issues was concerned with the explication of a conceptual model, with methodological procedures, in the area of time allocation in natural classroom situations. It has become increasingly evident to us that instructional time is related to educational outcomes. The BTES staff is continuing to investigate this area. The second methodological issue was concerned with an inquiry into generalizability theory applied to the problem of how many observers and/or how many occasions are necessary before stable estimates of a teacher's classroom behavior are recorded.

The goal of these five tasks, as well as some others, was to provide information that could be combined with existing data obtained by Educational Testing Service during 1973-1974, in order to design a large field study examining teacher effectiveness. As the planning for further research continued, the number of problem areas faced by the staff rapidly increased. While not as dramatic as solutions, the uncovering of problems is the first step toward doing competent research which can solve practical problems. This paper enumerates some of the problem areas faced by research workers concerned with teacher effectiveness. The paper was presented, in modified form, to the meetings of the National Association for Research in Science Teaching, Los Angeles, California, March 18, 1975. The author gratefully acknowledges the editorial contributions and ideas of Margaret Bierly, Leonard Cahen, Nikki Filby, Charles Fisher, Robert Heath, Richard Marliave, Marjorie Powell, and Richard Shavelson.

TABLE OF CONTENTS


| | |
|---|---|
| PREFACE | 1 |
| INTRODUCTION | 1 |
| INSTRUMENTATION PROBLEMS | 3 |
| Dependent Variable Problems | 3 |
| Standardized testing | 4 |
| Tests for special teaching units | 5 |
| Multivariate outcomes | 6 |
| Independent Variable Problems | 7 |
| Appropriateness of teacher behavior | 7 |
| The unit of analysis for the independent variable | 9 |
| Stability of teacher behavior | 10 |
| METHODOLOGICAL PROBLEMS | 13 |
| Student Background and Teacher Effectiveness | 14 |
| Subject Matter and Teacher Effectiveness | 15 |
| Normative Standards and Volunteer Samples in the Study of Teacher Effectiveness | 16 |
| Individual Differences Among Students and Teacher Effectiveness | 17 |
| Mediation of Teacher Effectiveness Through the Student's Behavior | 19 |
| Construct Validation and Teacher Effectiveness | 20 |
| The Generalizability of Measures of Effectiveness | 22 |
| STATISTICAL PROBLEMS | 23 |
| CONCLUSION | 25 |
| REFERENCES | 27 |

# IMPEDIMENTS TO THE STUDY OF TEACHER EFFECTIVENESS*

David C. Berliner
Far West Laboratory for Educational Research and Development

## INTRODUCTION

Advocates of performance or competency based teacher education, state mandated evaluation programs such as the Stull Bill in California, and teacher accountability systems, all suffer to some degree from ostrichism. Ostrichism is a common disease often afflicting education. It's etiology is in a premature commitment to a particular educational movement. Behavioral symptoms include the practice of sticking one's head into the sand when problems appear, in the hope that the problems will go away.

The particular educational movement which is inducing the current epidemic of ostrichism is the commitment of educators to competency training and evaluation without the existence of empirical evidence linking teacher behavior to student outcomes in classroom settings. The Coleman report (1966), and its offshoots (Jenks, 1972; Mosteller and Moynihan, 1972), have minimized the role of the teacher in accounting for educational outcomes. These investigators claim that family background, socioeconomic status, ethnicity and the like, are the major causal variables affecting between school differences in achievement. They imply that teachers only minimally

*The ideas presented in this paper have emerged from discussions with the staff of the Beginning Teacher Evaluation Study of the Far West Laboratory for Educational Research and Development. This is a project of the California Commission on Teacher Preparation and Licensing, funded by the National Institute of Education.

affect student achievement. Heath and Nielson (1974) reached the same conclusion in their review of the studies of teacher clarity, use of student ideas, criticism, enthusiasm, and other variables commonly accepted as skills or competencies. They concluded first that there is no established empirical relation between teacher behavior and student achievement. Second, that the flaws in the research are due to nonsensical statistical analyses, weak research designs, and sterile operational definitions of teacher behavior and student outcomes. And third, that because of the strong association between omnibus measures of student achievement and socioeconomic and ethnic status, the effects of teachers and techniques of teaching on achievement are bound to be trivial.

These are serious criticisms of the effects of teaching on student achievement. They have serious consequences since the heart of the performance and competency based approaches to teacher education, teacher evaluation and teacher accountability has to be the empirically established relationship between teacher behavior as an independent variable and student cognitive and affective outcomes as dependent variables. Whether we are interested in effective social or natural science teaching, or effective mathematics or home economics teaching, establishing empirical relationships between teacher behavior and student outcomes has to be our goal. Unless replicable findings relating teaching behavior to student achievement in natural classroom settings can be found, the performance and competency based teacher education, evaluation, and accountability programs will not be believable.

Ferment exists because performance and competency based education, in all its forms, has been sold before it really exists (cf. Shanker, 1974). Those who use research to criticize teachers, teaching, and performance

based teacher education, as well as those who defend teachers, teaching and performance based approaches have all taken positions before they have the necessary empirical backing. There is not now, and there will not be for some time any empirical evidence on which to take a firm position on these issues. Extremely important problems hamper the study of teachers and teaching in all subject matter areas. It will take years before these problems can even be understood well enough to do classroom research properly.

An important step in the systematic study of any phenomena is the recognition of what problems exist in that research area. Addressing these problems, rather than assuming they will go away, or that they do not apply, will enhance the likelihood that studies of teacher effectiveness will be fruitful. The problems can be loosely grouped into three categories concerned with the instrumentation, methodology, and statistics used in studying how teacher affect the achievement of students.

## INSTRUMENTATION PROBLEMS

There are serious instrumentation problems connected with both the independent and dependent variables commonly used in research on teacher effectiveness. Six of the problems are discussed here.

### Dependent Variable Problems

Our work at the Laboratory has been hampered by an inability to satisfactorily resolve three problems connected with development of dependent variables. These problems are connected with standardized testing, tests of special teaching units, and development of multivariate outcome measures.

Standardized testing. In studies of how teachers affect students, standardized achievement tests are extensively used as criteria or outcome measures. These tests are, as a group, highly reliable instruments. They usually have adequate curriculum content validity, and seem predictive of future academic success. These tests have, however, one overwhelming flaw. They simply may not reflect what was taught in any one teacher's classroom. The tests are designed to be used in all kinds of courses within a curriculum area, and therefore cannot be completely sensitive or appropriate for any one teacher's teaching (Gall, 1972). They simply lack content validity at the classroom level.

The standardized achievement tests are also highly correlated with standardized intelligence tests, thus causing us to wonder exactly what kinds of items are really used in these tests. Furthermore, the tests are usually group administered multiple-choice tests. When working with young, bilingual, or lower socioeconomic status children, there is a serious question about whether many of the children are being appropriately tested.

In our own work, when standardized tests must be used, we try to refine the items in a number of ways. We try to choose items where there is evidence of substantial change in difficulty level over some instructional period. In this way we hope to identify items that are reactive to instruction. We try to pick items that correlate weakly with a measure of general intelligence, like the Raven's Progressive Matrices test, rather than picking those items with higher saturations of general intelligence. We try to have teachers rate items on how much time it would take them to teach that idea, or, how much emphasis they put on material like that addressed by the item. Unless items on a standardized test are put through

a systematic screening of this type, the test is not going to be particularly reactive to teaching. Off-the-shelf standardized tests make poor dependent variables for studies of teaching. This is part of the difficulty in interpreting the Coleman report. The tests they used in that study were more reactive to family background and ethnicity than they were to instructional events within the school. It does not directly follow from this kind of evidence that teachers have no affect on student achievement.

Tests for special teaching units. To insure the use of tests that are content valid for a particular classroom, many investigators of teaching have created special teaching units, or content vehicles to study teaching (Berliner and Ward, 1974; Joyce, 1975; Popham, 1971). An experimental unit of this type contains curricula materials, objectives, and sample test items. The teacher is asked to teach to the objectives. The unit could be a single 30-minute lesson, or require daily work over three weeks. Under these conditions every teacher has similar materials and objectives to work with. Students are pretested and posttested with carefully constructed tests designed to tap many dimensions of the material in the experimental teaching unit. The dependent variable in this situation is much more valid and much more reactive to classroom teaching. In comparative studies of teaching effectiveness, these experimental teaching units, and their tests, have much to commend them. Each teacher has a similar chance to try to produce gains in student achievement. Some teachers will be better at this than others.

Unfortunately, at this time in our research efforts, we do not know if the measures of teaching effectiveness arrived at over a short period of time provide an estimate of teacher effectiveness over a longer period of time. This methodology, which is used in our research on teaching, allows us

to use tests of high content validity that seem to accurately reflect classroom practice for a short period of time. But this methodology may not always show strong predictive validity. The ranking of teachers on effectiveness, as determined by the relationships between student pretest and posttest scores associated with an experimental teaching unit, is only moderately correlated with a ranking of those teachers based on gains over the whole school year. Studying teacher effectiveness with dependent measures tied to special teaching units may not be a fair characterization of teaching over the long haul. But it certainly may be one way to identify teachers who differ in measured effectiveness when teaching a common curriculum, to common objectives, for controlled amounts of time.

Multivariate outcomes. There are at least two dependent variables in any instructional interaction that should be of interest to us. One of these is the achievement of the learner in the situation. This has been a commonly used measure of instructional outcomes. The other, less often examined, is the learner's feelings about the instructional situation. Students are not always asked questions which probe their liking for their teacher or the subject matter. Researchers often overlook inquiring about a student's enjoyment of their classmates, the degree of threat felt in the class, and whether or not they would take more courses in that area. Moreover, when such issues are addressed in research studies, the affective set of dependent measures is kept separate from the achievement measures.

The problem in the research we do is to find ways to use multivariate outcomes so that many kinds of achievement and affective responses are used as indicators of the quality of classroom life for a child. The problem is something like the difficulties in teaching reading. You can get

high comprehension at slow reading rates. Or you can get low comprehension at high rates of reading. But it is obvious that there must be some optimum multivariate outcome that simultaneously considers both reading comprehension and speed. The same kind of multivariate outcome measures, simultaneously considering both achievement and affective outcomes is needed for research on teaching. If researchers in this area do not consider what is learned and what is felt about that learning, simultaneously, they will continue to fractionate school learning into pieces that do not resemble the students' view of reality.

Independent Variable Problems

Research has also been hampered by problems connected with the independent variables used in studies of teacher effectiveness. A major difficulty is the "appropriateness" of particular teacher behavior in a given situation. A second issue is the determination of a unit of analysis for the independent variable. A third issue is the stability of teacher behavior.

Appropriateness of teacher behavior. Researchers have spent a good deal of time counting teacher behaviors. We know something about the number of higher and lower cognitive questions asked per unit time, we have counted the rate of positive verbal praise, the number of criticisms made, the number of probes, the frequency of explaining links, etc. For many of these variables a low correlation with some student outcome measure is found. But in classroom observation one becomes acutely aware of the difference between a higher cognitive question asked after a train of thought is running out, and the same type of question asked after a series of lower cognitive questions has been used to establish a foundation from which to explore

higher-order ideas. Teachers sometimes ask inane questions. Teachers sometimes direct questions to what we believe was the wrong child. We have seen positive verbal reinforcement used with a new child in the class, one who was trying to win peer group acceptance, and whose behavior the teacher chose to use as a standard of excellence. We watched silently as the class rejected the intruder, while the teacher's count in the verbal praise category went up and up and up. Teachers have been seen responding to student initiated questions with irrelevant information. Teachers sometimes achieve a high rate of probing student responses to questions, seemingly without regard for the student or the kind of initial response given to a question. Some students are embarassed by the probing, with other students probes occurred at inappropriate times, and sometimes probes were not used when the situation seemed to cry out for them. Similarly, skillful probing has been observed. A student's knowledge about an issue was brought out and shared with the class, after a weak first response was given by that student. The teacher's probing questions may have been as skillful as Plato's, but only their frequency was recorded.

All these events have led us to reassess our strong behavioristic stance in the study of teaching. We still regard frequency counts as very useful information. But we now feel quite strongly that the qualitative dimension, dealing with value judgments about appropriate use of skills, must enter into our observations of teaching. Researchers must address the appropriateness issue in order to study the information processing and decision making skills of human teachers. It is precisely these skills that provide the most important rationale for having human teachers in the classroom.

The unit of analysis for the independent variable. Something else we have become acutely aware of in our studies of teacher effectiveness is the problem of the unit of analysis for characterizing the independent variable. Is the single teacher question the unit of interest? Is the question, along with the wait-time, the unit? Or is the teacher question, wait-time, and student answer the unit which best characterizes the independent variable? And if the latter is most appropriate, does that transaction become part of an episode or strategy of even more complex dimensions and longer duration? Teachers follow strategies of questioning and of discussion. In an inductive lesson the meaningful unit of analysis may be a one-hour or one-week episode that is concerned with the conservation of matter. The individual questions, reinforcers, probes and student responses may be trivial aspects of the overall episode. New conceptions for the units underlying independent variables used in studies of teacher effectiveness are clearly needed.

Something else about the nature of an instructional episode is perplexing. Very little data is available describing the nature of the instructional activities and episodes a child engages in each day. Since instructional time appears to be an important variable in the learning process (Wiley, 1973; Harnischfeger and Wiley, 1975) accurate records of how time has been allocated to the various instructionnal activities and episodes is needed. The work of Gump (1967) and the techniques of Barker (1968) on obtaining accurate descriptions of the time a child spends in various activities may be useful starting points for obtaining these kind of data. The time and type of activities can be treated as independent variables and may be causally related to various types of student outcomes.

Stability of teacher behavior. Before an observer enters a classroom to code teacher behavior in any sensible way, he has to be sure of two things. First, that the event must occur frequently enough to observe during the observation period. Second, the behavior should be representative of the teachers usual and customary way of behaving. Only if these conditions are met can a teacher's behavior be sensibly characterized by the frequency count or rating scale description obtained in observation of classroom activities. These basic requirements for observation must be examined closely.

Many studies relating teacher behavior to student outcome have examined teacher behavior that did not occur frequently. For example, among 32 primary-grade science teachers the use of questions calling for identifying relationships, hypothesizing, and testing hypotheses is an extremely rare event on any given occasion of observation (cf. Moon, 1969; 1971). Another case of low frequency events, in an important area of teaching, has to do with the management skills of teachers. In some communities classroom management is not too difficult. The students are motivated and parents exert tight behavioral control at home, so that traumatic disturbances in classrooms are quite infrequent. In other communities serious management problems exist all day long. So we find that to observe instances of teacher behavior in the area of classroom management, ecological factors must be taken into account. Furthermore, even in settings where management problems usually occur with high frequency, certain teachers are so quick to establish a non-disruptive social system, that by the time the observer enters the class, particular kinds of events have been precluded from occuring.

How then can one study teacher behavior when important variables in the study rarely occur? One answer, of course, is in denser observation than is customary. Five one-hour observations of teacher behavior, which is unusually high for most studies of teaching, may simply not provide all the information an investigator may want. In addition, part of the answer is in knowing when and where to observe. For example, the first two weeks of schooling would be important for a study of management skills in inner city schools. Simply trying for denser observation later in the year and in other types of schools might be wasted effort.

The problem of estimating behavioral stability is partly related to the problem of the frequency of occurrence of behavior. When the frequency of a behavior is low the correlations between the frequency of occurrence for certain events, over occasions (that is, a coefficient of stability for the behavior), will be low. But part of the problem in looking at stability of teacher behavior is quite distinct from the frequency issue. Think for a moment about the characteristics you prize in a teacher. Usually, people think of "good" teachers as flexible. Such teachers are expected to change methods, techniques, and styles to suit particular students, curriculum areas, time of day or year, etc. That is, the standard of excellence in teaching commonly held implies a teacher whose behavior is inherently unstable. Needless to say, this poses a problem for an observer trying to observe a teacher's customary and usual ways of teaching.

For our study of teaching we have reviewed teacher stability, over occasions, for a great many variables (Shavelson and Dempsey, 1975). The results are fascinating. On the laughable side are coefficients of stability from Campell's (1972) analysis of science teaching at the junior high school level, over two occasions. The Flanders Interaction Analysis

System was used. The stability coefficient (that is, the correlation between a teacher's standing on a measure across two occasions) was -.50 for a measure of indirectness in teaching (the i/d ratio). On five occasions Moon (1969; 1971) studied 32 primary grade science teachers trained in the Science Curriculum Improvement Study (SCIS). The stability coefficient for the Flanders indirectness measure went all the way up to +.18; for the frequency of fact or recall questions, the stability coefficient was -.12; and for amount of teacher talk, only +.12. In Borg"s (1972) study, the behavioral stability of teachers was measured after training in questioning techniques had taken place. The stability of the ratio of higher-order to fact questions was .07. The rather large number of low and even negative stability coefficients which exist in the literature confirms our belief that the independent variables we often work with in studies of teacher effectiveness are not fair indicators of a teacher's typical behavior. Researchers are so eager to capture variables for data analysis with rating scales and frequency counts, that they seem to have forgotten to check if their methodology is appropriate to the phenomena they are interested in studying!

Of course there are many exceptions to the trend for teacher behavior to be unstable. We have found ratings of variables over 10 occasions that yield high stability coefficients. These include stability coefficients of .92 for teacher warmth; .79 for teacher enthusiasm; and .83 for teacher sensitivity (Wallen, 1969). We have found frequency counts demonstrating that a global variable composed of all types of reinforcement is reasonably stable over occasions, yielding a stability coefficient of .64 (Trinchero, 1974). In the latter study, however, there is considerable evidence pointing

to the lack of generalizability of stability coefficients across different teacher populations, curricula areas and student populations. For example, the stability coefficient over two occasions for the frequency of positive verbal teacher behavior was .04 for English teachers, and .57 for social studies teachers.

By examining the stability of teachers' behavior which is used as the independent variable in studies of teacher effectiveness, we conclude that: 1) some teacher behaviors that we think are important to study occur infrequently. To study them requires extensive observation in particular settings at appropriate times; 2) some teacher behaviors that we think are important to study are basically unstable over occasions. No practical amount of observation will result in a reliable estimate of a teacher's use of these behaviors. Perhaps we need to develop measures of variance instead of measures of central tendency to describe these behaviors; 3) some teacher behaviors are stable over occasions. In general, but not always, ratings or high inference variables, rather than frequency counts or low inference variables, are the more stable; 4) stability coefficients for many teacher behaviors will not demonstrate ecological or population validity. Teacher behavior is moderated, as it should be, by the kinds of students and the variety of settings that teachers work in. Until more is known about which teacher behaviors fluctuate, and how and why they fluctuate over time, settings, curricula, and populations, studies relating teacher behavior to student outcomes must remain primitive.

## METHODOLOGICAL PROBLEMS

A loosely related set of issues has been grouped under the title problems in methodology. Each of the problems and issues mentioned is in

some way hampering the development of reliable knowledge about the relationship between teacher behavior and student outcomes.

Student Background and Teacher Effectiveness

One problem in studying the teaching process is estimating how much can legitimately be expected of teachers or schools as an influence on student growth. This problem is debated in educational philosophy, sociology and economics, as well as educational psychology. And this issue has already been mentioned when it was noted that procedures were needed to reduce the influence of intelligence and ethnicity on test performance in studies of teacher effectiveness. But the problem is even more pervasive. Can a teacher be held accountable if a perfectly appropriate prescription is given, and then not followed by students? Suppose a teacher says, "read this chapter and come to my office so we can discuss it." Among sub-cultures that see schools as hostile or useless, students will not read the chapter and will not come in to discuss it. Classes of such students may show minimum growth in achievement at the end of the year. And these low achieving classes may very well be made up of lower socioeconomic status children and ethnic minorities. Under these conditions, how much responsibility is to be placed on teachers for the low student performance?

On the other hand, with children of high intelligence and high socioeconomic background, growth in achievement takes place almost in spite of teachers and teaching. Can the achievement of students in those settings be attributable to teachers, or is it a product of genetic and environmental advantage, relatively unaffected by what teachers do?

Since some children, often whole groups of children, may be unwilling to learn in the institutions now used to educate them, and some children learn in those institutions regardless of what happens to them, how do we go about attributing student achievement to what teachers do? In the case of low achieving students teachers may have to be evaluated against some other criteria than student achievement. Yet to do so denies that teachers can and should make a difference in the achievement of lower socioeconomic and minority children. There may not be solutions to this problem. But the problem exists and must be thought about as people naively discuss teacher effectiveness without qualifying what they say by noting the students' background characteristics, particularly socio-economic status and intelligence.

Subject Matter and Teacher Effectiveness

That student background characteristics influence test performance and almost all other aspects of schooling is well established. What was not so well understood, until recently, is that student performance in different curriculum areas is differentially affected by those background characteristics. In the International Education Association's (IEA) cross-cultural study of student achievement (Postlethwaite, 1973), the variance accounted for by student background characteristics, such as intelligence and social class, was estimated for a number of subject matter areas. Clearly highlighted, around the world, was that home influences on subjects like reading and social studies are very powerful. Those influences are so powerful in accounting for student achievement that there may not be enough variance unaccounted for in the performance of students to attribute to the influence of teachers.

But in other curriculum areas, student background accounts for much less variance. Physics, chemistry, French, Spanish, geometry, and trigonometry are not typically learned at home, and therefore the schools account

for more variance in these measures of achievement than for achievement measures in reading, social studies or language arts. This does not mean that socioeconomic status and intelligence are not related to performance in science, foreign language or mathematics. It simply means that the influence of those background factors is much less, thus leaving more variance to potentially attribute to school and teacher effects.

If teaching behavior in natural settings is to be studied in a correlational manner it should be studied in those areas of the curriculum where we are most likely to be able to attribute an effect to teachers, after the influences of test unreliability and home background have been removed. Instead researchers typically study teaching in those subject areas where they will be hardest pressed to causally relate teaching behavior to student outcomes. New approaches are called for.

Normative Standards and Volunteer Samples in the Study of Teacher Effectiveness

Much of the research on teacher effectiveness is, in simplest form, a comparison of the post-instruction test scores of classes that had similar pre-instruction test scores. These comparative differences in outcomes are believed to discriminate between more and less effective teachers. This research approach is entirely normative. And in a norm referenced research study some teachers will always appear to be better than others. In fact, the whole sample of teachers in any study may be quite poor when judged against some absolute standards, and we would never know.

More likely, since studies of teacher effectiveness in natural environments require the informed consent of volunteer teachers, research is likely

to be conducted with a sample of self-confident, relatively open teachers, almost all of whom may be superior to a non-volunteer sample on an unknown number of unidentified dimensions. But in a norm referenced system, where teachers are evaluated against other teachers, some of our sample will be judged to be less effective than others. This is a silly research strategy, but one that is not easily changed. To bring about change in this approach we would need to impose criterion referenced achievement standards for teachers, and require all teachers to participate in research on teacher effectiveness. Until that can be done (though I doubt it ever will be done) we should never talk of effective and noneffective teachers. At best the research sample can be described as more and less effective teachers, which is quite different from the absolute criteria implied by the terms effective and noneffective. And because norm referenced research is conducted with volunteer samples, statements about teacher effectiveness should also include some reference to the fact that these are more or less effective teachers from a sample of teachers that are themselves probably superior to the average teacher in an unknown number of ways.

Individual Differences Among Students and Teacher Effectiveness

Each teacher knows that some of the things they do will not be effective with some of the children they teach. There is no feeling of failure when this occurs, that's just the way things are. Most teachers recognize this problem and modify instruction accordingly. They customize their behavior, as best they can, to fit the individual styles of students. Research on teacher effectiveness, however, usually ignores this phenomena. Rarely is enough data about individual differences among students collected to find out if particular teaching behaviors are differentially effective with

different types of children. For example, from what is known about how aptitudes and treatments interact (cf. Berliner and Canen, 1973), it can be expected that a highly structured course in, say, science, taught by a well organized somewhat dominant teacher, will yield greater achievement for high anxious students than for low anxious students. On the other hand, the low anxious student will probably perform better than the high anxious student in the class of a science teacher providing only small amounts of guidance and using an inductive approach. Research on teacher effectiveness ordinarily finds no relation between student achievement and teacher behaviors that help to define constructs like inductive or deductive teaching style. Relationships may not appear because it is not yet known how to partition students into meaningful sub-groups for whom the two different treatments might be uniquely applicable. If students could have been divided into high and low anxious individuals, to follow our example, it might have been found that teacher behaviors within each teaching style had important effects on student achievement.

I have no doubt that the styles of teaching and teaching behavior recommended by, say, the curriculum guides accompanying new science curriculum projects are appropriate recommendations for some teachers, when interacting with some students. But not all students! By not focusing on the individual aptitudes, styles, personality, and traits of the students, the effects of teachers are masked thus making it almost impossible to establish empirical relations between teaching behavior and student outcome.

An equally important reason to use the aptitude-treatment interaction approach is to find teacher behaviors that in general have positive relationships with student outcomes, but are, in fact, negatively affecting the

performance of small numbers of students. Research on teacher effectiveness has to begin searching for interactions as it continues trying to establish more general links between teacher behavior and student outcomes.

Mediation of Teacher Effectiveness Through the Student's Behavior

A fact of classroom reality that must be brought into designs for research on teaching is that teacher behavior does not influence student achievement directly. A teacher's indirectness, or questioning, or reinforcement does not simply result in greater mathematics, reading, or science achievement. The link that must be considered is the behavior of the student in the instructional setting. We are now convinced that the mediating link so necessary to consider is a student's active time-on-task. If teacher questions, reinforcement, warmth, and clarity are to affect outcomes, they can only do so by engaging and then keeping the student's attention. If the student will attend, the possibility of learning exists. Teacher behaviors that affect student active learning time must be examined carefully. To do so means putting much more effort into clinical studies. In this way an investigator can work one-to-one with students, trying to understand how the student allocates his attention, and how nominal stimuli emitted by the teacher, become effective stimuli for that student. To think that there is a direct link between, say, a teacher's questions which require the generation of hypotheses by students, and the student's achievement on an achievement test is overly simple. Intermediate links in that causal flow require us to examine the student's attending and information processing behavior.

Another aspect of the student that must be thought about for research in teaching is the student's perspective of the events that impinge upon

him in classrooms. Researchers do not know how much of what is called skilled teaching is even perceived by the learner. From the learner's perspective, perhaps "analysis" and "synthesis" level questions are not distinguishable. Students may differentiate only "memory" and "thinking" questions. From the learner's perspective the rate of reinforcement may be irrelevant. The teacher either is "nice" or "not nice" to students. I believe that some variables thought to be quite important by educational theorists are in fact unimportant, unperceived, or unperceivable by students (cf. Winne, 1974). Students exposed to variables they cannot perceive or to variables they believe to be unimportant, may be unaffected by such variables. Researchers certainly need to follow Snow's (1974) advice to researchers that urges more detailed accounts of what learners do in response to experimental treatments.

Construct Validation and Teacher Effectiveness

Through the writings of the logical positivists, and particularly the physicist Bridgeman, social scientists became aware of the critical nature of language and operations in science. An initial development to further scientific understanding of some phenomena is a descriptive language that uses concepts having common meaning among the scientists working in the same area. The intensive and extensive meaning of key concepts needs to be shared by the members of the scientific community. The less the overlap of shared meaning, the less rigor the science can develop. A case in point would be a term like "withitness" from the study of teaching by Jacob Kounin (1970). The teacher who can spot trouble before it begins has "withitness." Such a teacher can be working with one group of students and call out a student's name at the other end of the room because he is beginning to cause a dis-

turbance. That is "withitness". I recently went into a classroom and one of the concepts that helped me organize what I saw was the concept of "withitness". I felt perfectly at home using the concept. It helped me make sense out of the different styles of two teachers I was observing. Yet the concept itself cannot be rigorously defined and relies upon very subjective interpretation of phenomena. The construct of "withitness", like many of the concepts we work with, is useful, but inadequately defined.

One way to increase the preciseness of our concepts is to tie them through clear operations to the measurement of their occurrence. For example, a concept like teacher warmth can be defined as the number of times per day the teacher smiles. But is that what is wanted when warmth is to be measured? It seems that the phenomena of interest is fragmented beyond recognition when the occurrence of some molecular behavior is used to operationally define our terms.

What is needed in the study of teaching is to begin incorporating multiple methods of measurement into the studies we do (Campbell and Fiske, 1959). If one chooses to work with the concept of "withitness" or "warmth", there is a need to measure the concept from as many different perspectives as we can. For example, a teacher's warmth can be measured by self-report, student report, observer rating, frequency count of smiles, percent of gestures regarded as affectionate, and anything else that can be thought of. Then, from the intercorrelations of the various imprecise and imperfect measures of warmth, one can begin to understand the construct that is so glibly used, but cannot clearly be defined. Extensive construct validation must take place or the impreciseness of our language for describing the phenomena of interest will keep the empirical study of teaching at its present primitive level.

The Generalizability of Measures of Effectiveness

If teachers are to be characterized as more or less effective, in order to see if the behavior of those teachers differ, knowledge about whether the teachers maintain their rank ordering on measures of effectiveness over time and over subject matter areas is needed. As part of our research we reviewed studies that addressed this problem. There are about eight studies of teacher effectiveness over lengthy periods of time (see Shavelson and Dempsey, 1975). From those studies it is estimated that the mean correlation between measures of teacher effectiveness obtained two or more times is about .30. This estimate is based on data from predominantly primary age children tested with standardized reading and mathematics achievement tests. Brophy's (1973) study presents some interesting data to consider. Residual gain scores over 3 years were examined for 165 elementary teachers. Twenty-eight percent of the teachers were consistent in their effects on students three years in a row. Approximately 14 percent of the teachers in the study were consistently effective in producing higher than predicted reading and math achievement. And 14 percent of the teachers were consistent in being associated with classes that had scores lower than predicted in reading and mathematics three years in a row. Thirteen percent of the teachers showed linear increases in residual gains over the three years. That is, they appeared to be getting more effective in their teaching. Similarly, 11 percent of the teachers showed a linear decrease over that time period. They seemed to be getting less effective over time. Forty-nine percent of the teachers in this sample were inconsistent in the patterning of their residual scores over time.

In our review of short term studies of teacher effectivenss, ranging

across grade levels and all kinds of curriculum areas, it was found that when the same content is taught to similar students (for example, teaching and reteaching an ecology lesson to two samples of urban students), moderately stable estimates of teacher effectiveness are obtained. But when different content is taught to two or more groups of similar students, the effectiveness measures were not stable. Similarly, when different content is taught to the same students, effectiveness from occasion to occasion was unstable. In recent research, involving 200 elementary school teachers, each of which taught a two week, specially designed teaching unit in reading and mathematics similar data was obtained. Residual gain scores for each subject matter were calculated. These measures of effectiveness using different content and the same students were correlated. From these data we find that measures of effectiveness in the two curriculum areas correlate about .30.

It appears that teachers do not, by and large, remain in a stable ordering on measures of teacher effectiveness. If, as has been discussed, the independent variables typically looked at are often unstable, and measures of teacher effectiveness also show instability, the possibility of correlating teacher behavior with student achievement to determine effective teaching behavior is quite limited. In fact, unless we reconceptualize much of what we do in this research area, our research will be ludicrous!

## STATISTICAL PROBLEMS

Instrumentation and methodological problems have been discussed above. A brief discussion of the statistical problems associated with the study of teacher effectiveness is in order. One strategy that has been used in our research is to identify groups of teachers that differ in effectiveness and

then to analyze the teaching behavior of teachers in the contrasting groups. Our choice of statistical techniques is limited to those that apply when a single achievement test is administered to students prior to, and following some teaching; and the teaching is considered an intervention that takes place with students who were not randomly assigned to classes. Under these conditions a statistical method is required to discriminate between groups of teachers that differ significantly in average pupil gain. The basic problem is one addressed over and over in educational research. How do you measure change without a true experimental design?

The whole range of statistical techniques based on regression approaches has been examined. The advantages and disadvantages of residualized true scores, curvilinear adjustments and methods that correct for non-homoscedastic bivariate distributions have been analyzed. Ways to define effectiveness based simply on posttest raw score differences, for classes that had similar pretest scores have also been explored. And there is much to recommend in this simplest of methods, which avoids all pretense of sophisticated statistics. There are interesting possibilities in the new scaling methods, which avoid many of the assumptions of classical test theory. Groups of teachers that maximally differ from each other can be identified with these techniques, providing samples of more and less effective teachers within curriculum areas.

But, in general, most researchers are now using statistical procedures in these areas which one can put little faith in. The procedures almost always do not adequately meet the requirements of the situation under study.

Sophisticated statisticians working in applied settings are needed to help in this kind of research work. But that will not happen given present funding priorities and limitations.

## CONCLUSION

Stated above was the belief that the heart of performance and competency based teacher education, evaluation and accountability programs is the establishment of empirical relationships between teacher behavior as an independent variable and student achievement as a dependent variable. But before researchers can adequately establish those relationships they need to deal with the problems of instrumentation, methodology and statistics. Workers in this area must come to grips with the inadequacy of standardized tests, the unknown predictive validity of tests from special teaching units, the problem of building multivariate outcome measures, the problems of measurement of appropriateness of teacher behavior, the lack of experience in choosing an appropriate unit of analysis for describing teaching behavior, and the lack of stability of many teacher behaviors.

Time must be taken to consider the problems of how student background affects measures of teacher effectiveness, what subject matters should be examined, how normative standards and volunteer teachers affect what can be said about teachers and teaching, how individual students react to teaching skills, and how students monitor and interpret a teacher's behavior in ways which may or may not coincide with how educational theorists interpret the phenomena. Time and resources are needed to do construct validation and studies of the generalizability of measures of teacher effectiveness.

Finally, guidance is needed for choosing techniques to use for measurement of change in the achievement of students in natural classrooms.

When we have finished examining this potpourri of problems, issues, and concerns, we will be ready to begin the scientific study of teaching. And if we cannot deal with all of these problems, perhaps we should simply acknowledge that teaching is, after all, a very complex set of events which cannot be easily understood.

## REFERENCES


Barker, R. G. (1968) Ecological psychology: Concepts and methods for studying the environment of human behavior. Stanford, California: Stanford University Press

Berliner, D. C. and Cahen, L. S. (1973) Trait-treatment interactions and learning. In F. N. Kerlinger (Ed.), Review of research in education, 1. Itaska, Illinois: F. E. Peacock Publishers.

Berliner, D. C. and Ward, B. A. (1974) Proposal for Phase III Beginning Teacher Evaluation Study. San Francisco, California: Far West Laboratory for Educational Research and Development

Borg, W. R. (1972) The minicourse as a vehicle for changing teacher behavior: A three-year follow-up. Journal of Educational Psychology, 63, 572-579

Brophy, J. E. (1973) Stability of teacher effectiveness. American Educational Research Journal, 10, 245-252.

Campbell, D. T. and Fiske, D. W. (1959) Convergent and discriminant validation by the multi-trait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Campbell, J. R. (1972) A longitudinal study in the stability of teachers' verbal behavior. Science Education, 56, (1), 89-96.

Coleman, J. S. et al (1966) Equality of educational opportunity. Washington, D. C.: U. S. Government Printing Office.

Gall, M. D. (1973) The problems of "Student Achievement" in research on teacher effects. San Francisco, California: Far West Laboratory for Educational Research and Development (Report A73-2).

Gump, P. V. (1967) The classroom behavior setting: Its nature and relation to student behavior. U. S. Office of Education, Dept. of Health, Education and Welfare. Final Report, Project No. 5-0334, Contract No. OE-4-10-107. Lawrence, Kansas: University of Kansas (mimeo).

Harnischfeger, A. and Wiley, D. E. (1975) Teaching-learning processes in elementary school: A synoptic view. Beginning Teacher Evaluation Study, Technical Report No. 75-3-1. San Francisco, California: Far West Laboratory for Educational Research and Development (mimeo).

Heath, R. W. and Nielson, M. A. (1974) Performance-based teacher education. Review of Educational Research, 44, 463-484.

Jencks, C. et al. (1972) Inequality: A reassessment of the effect of family and schooling in America. New York: Basic Books

Joyce, B. R. (1975) Vehicles for controlling content in the study of teaching. Paper given at the meeting of the American Educational Research Association, Washington, D. C., April, 1975.

Kounin, J. S. (1970) Discipline and group management in classrooms. New York: Holt, Rinehart and Winston.

Moon, T. C. (1969) A study of verbal behavior patterns in primary grade classrooms during science activities. Unpublished doctoral dissertation, Michigan State University, East Lansing, Michigan.

Moon, T. C. (1971) A study of verbal behavior patterns in primary grade classrooms during science activities. Journal of Research in Science Teaching, 8, 171-177.

Mosteller, F. and Moynihon, D. P. (1972) On equality of educational opportunity. New York: Vintage Books.

Popham, W. J. (1971) Performance tests of teaching proficiency: Rationale, development, and validation. American Educational Research Journal 8, 105-117.

Postlethwaite, T. N. (1973) A selection from the overall findings of the IEA study in science, reading comprehension, literature, French as a foreign language, English as a foreign language, and civic education. Paris, France: International Institute for Educational Planning (mimeo report no. IIEP/STU/MISC/73.3 [Rev. 1]).

Shanker, A. (1974) Competency-based teacher training and certification: Acceptable and unacceptable models. QuEST Consortium Yearbook. Washington, D.C.: American Federation of Teachers.

Shavelson, R. S. and Dempsey, N. (1975) Generalizability of measures of teacher effectiveness and teaching process. San Francisco, California: Beginning Teacher Evaluation Study, Technical Report #2, Far West Laboratory for Educational Research and Development.

Snow, R. E. (1974) Designs for research on teaching. Review of Educational Research, 44, 265-291.

Trinchero, R. L. (1974) Three technical skills of teaching: Their stability and effect on pupil attitudes and achievement. Unpublished doctoral dissertation, Stanford University, Stanford, California.

Wallen, N. E. (1969) Sausalito teacher education project. Annual Report. San Francisco, California: Division of Compensatory Education, Bureau of Professional Development, San Francisco State College (mimeo).

Wiley, D. E. (1973) Another hour, another day: Quantity of schooling, a potent path for policy. Studies of Educative Processes, No. 3, Chicago, Illinois: University of Chicago, July 1973.

Winne, P. H. (1974) Teacher effectiveness and student perceptions of teacher cues. Dissertation proposal, Stanford School of Education, Stanford, California (mimeo).